Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Combining structural equation modeling, necessary condition analysis and explainable machine learning reveals distinct drivers and preconditions of student ChatGPT adoption

Loi Huu Nguyen

loinh14@fe.edu.vn

FPT University

Anh Thai Nguyen

Van Lang University

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11278294/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Combining structural equation modeling, necessary condition analysis and explainable machine learning reveals distinct drivers and preconditions of student ChatGPT adoption

Nguyen Huu Loi [1*] and Nguyen Thai Anh [2]

[1*]FPT Polyschool, FPT University, Can Tho City, Vietnam.
[2]Faculty of Information Technology, Van Lang University, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): Loinh14@fe.edu.vn;
Contributing authors: anh.nt@vlu.edu.vn;

Abstract

Universities need to know which factors raise students' intention to use generative artificial intelligence and which must be present before high intention can occur. The hybrid structural equation and neural network analyses that dominate this literature address only the first question, rank predictors with a heuristic importance index and rarely test prediction on new respondents. We re-analyzed an open survey of 1,242 ChatGPT-using university students from 78 Philippine institutions, combining partial least squares structural equation modeling, necessary condition analysis, gradient boosting explained with Shapley values and cross-validated predictive tests. Habit was the strongest unique predictor of intention to use ChatGPT yet ranked sixth or seventh of seven on necessity under every scoring, whereas effort expectancy, with no detectable unique effect, ranked among the top three. Necessity magnitudes depended on the Likert response distributions and fell below conventional thresholds after rank transformation, so the robust result is this inverted ordering rather than the magnitudes. The three importance rankings agreed at the top and bottom, and neither the structural model nor the boosted ensemble predicted new respondents better than a linear benchmark. Separating predictors that raise intention from those that must be present gives institutions a clearer basis for sequencing support.

Keywords: ChatGPT, Technology acceptance, Necessary condition analysis, Structural equation modeling, Explainable artificial intelligence, Higher education

Introduction

Generative artificial intelligence (AI) has spread rapidly through higher education since ChatGPT was released in November 2022. In a mid-2024 international survey, 86% of students reported using AI in their studies, with ChatGPT the most used tool at 66% [1]; by early 2025 the share of United Kingdom undergraduates using AI in some form had reached 92%, up from 66% a year earlier, and the share using generative AI for assessments had risen from 53% to 88% [2]. Universities now have to decide where to invest in infrastructure, training and academic-integrity policy, and that decision rests on two different questions. The first is which factors raise students' intention to use tools such as ChatGPT. The second is which factors must be in place before high intention can occur at all. The first is a question of sufficiency and the second a question of necessity, and the evidence base addresses almost exclusively the first.

The dominant quantitative approach applies the extended unified theory of acceptance and use of technology (UTAUT2) [3, 4] with partial least squares structural equation modeling (PLS-SEM), frequently followed by an artificial neural network (ANN) in a hybrid SEM-ANN pipeline [5]. PLS-SEM tests linear hypotheses, and the ANN then re-ranks the predictors with a normalized importance index derived from sensitivity analysis or connection weights. Applied to ChatGPT, this work has converged on habit, performance expectancy and hedonic motivation as the strongest antecedents of students' intention [6–9], and configurational analyses have suggested that purely additive models miss part of the adoption story [10]. Building on a systematic review of the pipeline [5] and the wider methodological literature, we see three limitations. First, normalized importance is a heuristic quantity: it depends on network architecture and training runs, and unlike Shapley-value attributions [11, 12], which are beginning to appear in acceptance research [13], it has no axiomatic foundation. Second, the pipeline evaluates only sufficiency, the additive logic in which more of a predictor yields more of the outcome, although the same review recommends complementing it with necessity logic [5, 14, 15]. Third, studies often report the network's internal cross-validation rather than a genuine out-of-sample assessment of the structural model [16, 17].

Necessity is not a refinement of sufficiency but a different relation. A condition is necessary for an outcome when high levels of the outcome do not occur at low levels of the condition, whatever the other predictors do. In a scatterplot this appears as an empty upper-left corner, which necessary condition analysis (NCA) quantifies with a ceiling line [14, 18]. Because the two relations are logically independent, a predictor can have a large average effect yet be fully compensable by other factors, or a negligible average effect yet be indispensable [15]. Combined importance-performance map analysis builds decision support on exactly this pairing [19], and a European study that paired PLS-SEM with NCA identified performance expectancy among the necessary conditions for ChatGPT acceptance [20].

Here we address the three limitations together on a single openly licensed dataset, the item-level UTAUT2 survey of 1,242 ChatGPT-using university students from 78 Philippine institutions [21], which its creators analyzed with WarpPLS-based

PLS-SEM [22]. We use a four-stage framework (Fig. 1): PLS-SEM for measurement validation and hypothesis testing; NCA with permutation inference; a tuned gradient-boosting model (XGBoost [23]) interpreted with SHapley Additive exPlanations (SHAP) computed by TreeSHAP [12] and benchmarked against the conventional ANN and a linear model on a locked test set; and PLSpredict with the cross-validated predictive ability test (CVPAT) for out-of-sample assessment of the structural model [16, 17, 24].

The primary contribution is substantive. The sufficiency and necessity orderings of the UTAUT2 determinants agree in the middle but invert at the extremes. Habit, the strongest unique predictor of intention, ranks sixth or seventh of seven on necessity under every scoring we examined, whereas effort expectancy, which has no detectable unique effect once the other predictors are controlled, ranks among the top three; across the seven predictors the two orderings are essentially uncorrelated (Spearman ρ between -0.11 and 0.21 across three scorings). We display this pattern in a necessity-sufficiency map that applies combined importance-performance logic [19] to the sufficiency-necessity plane. Two further results frame the interpretation. A formal concordance test, which replaces the customary informal claim that rankings are aligned [25], shows that the attribution methods agree at the top and bottom of the sufficiency ranking, and a transparent out-of-sample assessment shows that neither the theory-constrained model nor a tuned ensemble predicts better than linear regression. Because the sufficiency analysis is thus stable at its extremes and offers no predictive gain over a linear model, the necessity analysis supplies the complementary information that the conventional pipeline lacks. We present the framework as an integration and replication-extension rather than as the first combination of any pair of methods: PLS-SEM has been combined with NCA for ChatGPT acceptance [20], structural equation modeling with Shapley-based machine learning for generative AI adoption [26], and SEM-ANN analyses with necessity logic for mobile sharing services [27] and, through combined importance-performance maps, for ChatGPT-4.0 use [28].

Results

Measurement and structural model

The sample was young and at an early stage of adoption: 95.5% of respondents were 25 or younger, 53.3% identified as female, and 81.2% placed themselves in the lowest self-rated ChatGPT experience band (Supplementary Table S1). The purification rule (Methods) removed two indicators, EE4 (loading 0.474) and BI1 (loading 0.668). BI1 was removed because the average variance extracted (AVE) of behavioral intention was 0.499 and rose to 0.542 after removal; the removal also raised the heterotrait-monotrait ratio (HTMT) between intention and use from 0.916 to 0.920, and retaining BI1 leaves the main necessity contrast unchanged (see Robustness). The rule's discriminant-validity safeguard retained the next-weakest effort expectancy item, EE3: removing it would have raised the construct's AVE to 0.569 but pushed the HTMT between performance and effort expectancy from 0.887 to 0.976. Composite reliability ranged from 0.796 to 0.900 and Dijkstra-Henseler ρ_A from 0.697 to 0.863. Effort expectancy

was marginal on Cronbach's alpha (0.672), ρ_A (0.697) and AVE (0.496), which we accept given its composite reliability (0.796) and the documented discriminant cost of further deletion (Supplementary Tables S2 and S3). The Fornell-Larcker criterion held for every pair of constructs. Two HTMT values exceeded 0.85, performance-effort expectancy (0.887) and intention-use (0.920); their one-sided 95% bootstrap upper bounds (0.937 and 0.951) exceed the stricter threshold of 0.90 but remain below 1 (Supplementary Tables S4 and S5), so discriminant validity for these two conceptually adjacent pairs is supported only under the liberal criterion [29, 30]. This is one reason we treat intention as the primary outcome. The first principal component of the items (Harman's single-factor test) explained 32.5% of item variance and full-collinearity variance inflation factors ranged from 1.76 to 2.68, so these diagnostics gave no clear signal of severe common method bias.

The model explained 44.6% of the variance in behavioral intention and 58.9% in use behavior (Table 1; Supplementary Fig. S1). Habit was the strongest antecedent of intention ($\beta = 0.293$, 95% confidence interval (CI) 0.239 to 0.346, $P < 0.001$, $f^2 = 0.105$), followed by facilitating conditions ($\beta = 0.170$, $P < 0.001$), hedonic motivation ($\beta = 0.166$, $P < 0.001$), performance expectancy ($\beta = 0.103$, $P = 0.003$) and price value ($\beta = 0.079$, $P = 0.034$). Effort expectancy ($\beta = 0.034$, $P = 0.300$) and social influence ($\beta = 0.003$, $P = 0.866$) had no detectable unique effect. Effort expectancy nevertheless correlates moderately with intention ($r = 0.46$), so its small path reflects its overlap with the other predictors, particularly performance expectancy. Intention ($\beta = 0.452$), habit ($\beta = 0.289$) and facilitating conditions ($\beta = 0.176$) all predicted use (each $P < 0.001$), so eight of the ten hypotheses were supported at $\alpha = 0.05$; under a Holm correction across the ten paths, the price value path no longer qualifies ($P = 0.034$ against a threshold of 0.017). Of 30 moderation terms for age, gender and experience, two had nominal $P < 0.05$ (age $\times$ effort expectancy, $\beta = -0.108$, $P = 0.006$; age $\times$ hedonic motivation, $\beta = 0.098$, $P = 0.024$), and neither survived Bonferroni correction ($\alpha = 0.0017$; Supplementary Table S6). The dataset creators reported moderating roles of age, gender and experience [22]; with heteroskedasticity-consistent standard errors and correction for multiple testing, these effects are not confirmed here.

Necessary conditions

Table 2 reports necessity effect sizes d under two ceiling techniques, and Fig. 2 shows four representative ceilings. We treat the ceiling-regression line (CR-FDH) on latent variable scores as primary and the step-wise ceiling-envelopment line (CE-FDH), which gave larger estimates for every pair, as an approximate upper bound (Methods). The necessity results diverged from the structural model. Effort expectancy, which had no detectable unique path to intention, had the largest necessity estimate ($d = 0.147$, $P < 0.001$), followed by facilitating conditions ($d = 0.118$, $P < 0.001$), hedonic motivation ($d = 0.115$, $P < 0.001$) and performance expectancy ($d = 0.110$, $P = 0.002$). Habit, the strongest sufficiency predictor, had only a small necessity effect ($d = 0.033$, $P = 0.002$), detectable but far below the conventional relevance threshold of 0.10, and even the upper bound of its upward-shifted bootstrap interval (0.091) remained below that threshold. Social influence ($d = 0.037$, $P = 0.205$) and price value ($d = 0.068$, $P =$

0.087) also fell short. The CE-FDH bottleneck analysis makes the contrast concrete: reaching 90% of the observed intention range required at least 64.8% of the observed range of effort expectancy but only 19.7% of the range of habit (Supplementary Tables S7 and S8). The bootstrap intervals of d are shifted upward, because each resample contains only a subset of the observed cases and its ceiling therefore lies on or below the full-sample ceiling, and they overlap within the cluster of necessary conditions, so they do not license a rank ordering within that cluster. The firm result is the contrast between the extremes of the necessity ordering, effort expectancy versus habit, and it holds on behavioral intention alone. For use behavior, intention ($d = 0.187$) and facilitating conditions ($d = 0.135$) were necessary, whereas habit again had only a small effect ($d = 0.042$, $P < 0.001$).

Judged against the necessity propositions (Methods), performance expectancy was conditionally supported as necessary for intention (NP1; borderline under construct-mean scores, see below), habit was rejected as necessary for use (NP2), and facilitating conditions were supported as necessary for use (NP3). The negative-control proposition (NP4) received split support. Price value behaved as intended, falling short under CR-FDH although CE-FDH gave $d = 0.107$ ($P = 0.004$), whereas hedonic motivation unexpectedly retained medium necessity under both ceilings. These results reproduce the necessity of performance expectancy reported for a European sample [20], whereas habit, the strongest sufficiency predictor here, is not a meaningful necessary condition.

Necessity-sufficiency map

Fig. 3 places each predictor of intention by its path coefficient (sufficiency) and its CR-FDH effect size (necessity), with reference lines at the conventional thresholds of 0.10. We call predictors in the necessary-but-not-sufficient quadrant *floors* and those in the sufficient-but-not-necessary quadrant *engines*. Habit is the engine: higher habit is associated with higher intention, but low habit does not preclude high intention. Effort expectancy is the floor, a candidate precondition rather than a lever. Facilitating conditions and, as boundary cases, hedonic motivation and performance expectancy are core preconditions that are both necessary and sufficient; performance expectancy lies within rounding of both lines ($\beta = 0.103$, $d = 0.110$), so its quadrant is fragile. Social influence and price value are peripheral; price value has a detectable but small path ($\beta = 0.079$, $P = 0.034$) that lies below the sufficiency line. The error bars show that the map is a qualitative partition. What is firm is that habit lies far to the lower right and effort expectancy to the upper left. This contrast is not an artifact of comparing partial path coefficients with bivariate ceilings: on zero-order correlations with intention, habit ranks second ($r = 0.54$) and effort expectancy sixth ($r = 0.46$), and the correlation ranking is unrelated to the necessity ranking (Spearman $\rho = 0.00$). A single importance ranking cannot express the distinction because it collapses both logics onto one scale.

Robustness of the necessity estimates

NCA on five-point Likert data is sensitive to the shape of each predictor's distribution, because a right-shifted, ceiling-heavy variable can empty the upper-left corner

of the scatter (raising d) while its restricted range lowers its path coefficient. Effort expectancy is such a variable: it has the highest mean, the lowest standard deviation and the largest share of respondents near the scale maximum of the seven predictors (Table 3). Across the seven predictors, d tended to increase with the predictor mean (Spearman $\rho = 0.75$, exact permutation $P = 0.066$) and to decrease with skewness ($\rho = -0.64$, $P = 0.139$), associations that are suggestive rather than conclusive with seven predictors. Performance expectancy and hedonic motivation crossed the 0.10 threshold under latent variable scores but not under construct-mean scores ($d = 0.092$ and 0.097), whereas effort expectancy (0.144) and facilitating conditions (0.119) remained above it under both. A rank transformation of the construct-mean scores, which equalizes all marginal distributions, reduced every effect below the threshold, including those of effort expectancy ($d = 0.036$) and facilitating conditions ($d = 0.055$), although permutation tests on the ranks still detected non-zero ceilings for six of the seven predictors ($P \leq 0.005$). Because necessity is defined on absolute levels, which the rank transform discards, this is a deliberately conservative bound on magnitude rather than evidence against necessity. The ordering was stable throughout: habit ranked seventh, sixth and sixth of seven on necessity under latent variable, construct-mean and rank scorings, and effort expectancy first, first and third.

Three further checks support the contrast (Supplementary Table S13). Deleting any single ceiling-defining case left effort expectancy (0.140 to 0.158) and facilitating conditions (0.119 to 0.153) above the threshold and habit (0.023 to 0.053) below it, whereas the classification of the other four predictors depended on single cases (for example, performance expectancy 0.090 to 0.150). Ceilings estimated on the development partition were exceeded by at most one of the 187 test cases for every predictor, so the ceilings generalize to unseen respondents. Retaining BI1 left the contrast intact (effort expectancy $d = 0.138$, habit $d = 0.049$ with $\beta = 0.306$) and made performance expectancy a clearer core precondition ($\beta = 0.160$, $d = 0.150$). The specific necessity magnitudes, and the finer floor-versus-core distinctions, are therefore properties of the observed Likert operationalization that cross-sectional data cannot separate from distributional structure, whereas the inverted ordering at the extremes is robust.

Because habit's lack of necessity contrasts with the European evidence [20], we asked whether necessity differs between self-rated novices ($n = 1{,}009$) and more experienced users ($n = 233$) (Supplementary Fig. S2). Several raw estimates were larger among experienced users, for example for performance expectancy (CR-FDH $d = 0.096$ for novices and 0.188 for experienced users). Necessity effect sizes, however, tend to grow as samples shrink, because fewer cases are available to fill the upper-left corner. Against 1,000 random novice subsamples of the experienced group's size, every experienced-group estimate lay within the central 95% of the size-matched distribution, and ten of twelve lay below its median (Supplementary Table S13). The data therefore give no evidence that necessity changes with experience, and habit did not reach the threshold for intention in either group.

Out-of-sample prediction

All nine endogenous indicators achieved positive Q^2_{predict} (0.208 to 0.379; Supplementary Table S9), so the structural model clearly outperformed the naive indicator-average benchmark, as CVPAT confirmed (overall average loss difference -0.263, $P < 0.001$; Table 4a). Against indicator-wise linear regressions the model did worse. It attained a lower root mean squared error (RMSE) than the linear benchmark on only two of the nine indicators, and the CVPAT loss difference, for which positive values favor the benchmark, was indistinguishable from zero for intention (-0.002, $P = 0.424$) but positive for use behavior ($+0.018$, $P < 0.001$) and for all outcomes pooled ($+0.009$, $P < 0.001$). In the vocabulary of the PLSpredict guidelines [16], the model has low predictive power relative to a linear benchmark.

The machine learning comparison pointed the same way (Table 4b). Cross-validated RMSEs on the development partition were nearly identical (0.552 to 0.555 for intention). On the locked test set ($n = 187$), the tuned XGBoost model attained the lowest intention RMSE (0.573, $R^2 = 0.392$) against 0.589 for linear regression and 0.591 for the ANN, but a paired bootstrap of the test cases put its difference from linear regression at -0.016 (95% CI -0.032 to 0.001; paired t test on squared errors, $P = 0.071$), and XGBoost had the highest cross-validated RMSE of the three models. For use behavior, the three models were within 0.005 RMSE of one another. Random search selected a shallow, heavily regularized ensemble (tree depth 2, with the largest grid values of the L1 penalty, the split-loss penalty and the minimum child weight) and the smallest, most strongly penalized network (two hidden units), indicating a response surface close to additive.

Agreement between importance rankings

Fig. 4 compares the three importance attributions for intention. Habit had the largest mean absolute SHAP value (0.164), followed by hedonic motivation (0.125) and facilitating conditions (0.096), with performance expectancy (0.038), effort expectancy (0.034), social influence (0.024) and price value (0.021) far behind. The strongest pairwise SHAP interaction, habit $\times$ hedonic motivation (0.022), was about one-eighth of habit's mean absolute SHAP value on the same full sample (0.178; Supplementary Table S10), consistent with the shallow, near-additive ensemble selected by tuning. The three rankings agreed more than random rankings would (Kendall's $W = 0.881$, permutation $P < 0.001$; chi-square approximation $\chi^2(6) = 15.86$, $P = 0.015$), a low bar for three methods fitted to the same data, so the bootstrap intervals of pairwise agreement are the more informative evidence. Pairwise Spearman correlations were 0.857 (SHAP versus path coefficients), 0.750 (SHAP versus ANN) and 0.857 (path coefficients versus ANN), with Kendall's τ between 0.62 and 0.71 and bootstrap lower bounds for τ between 0.43 and 0.62 (Supplementary Table S11). Agreement was concentrated at the extremes of the ranking and disagreement in the middle ranks, where incremental advice is usually drawn: price value ranked seventh by SHAP, fifth by path coefficient and fourth by ANN importance, and performance expectancy moved between fourth and sixth (Supplementary Table S12). For use behavior, all three methods produced the same ordering (intention, habit, facilitating conditions).

Discussion

The sufficiency results replicate the emerging consensus on ChatGPT adoption: habit dominates intention, intention and habit dominate use, and effort expectancy and social influence lose their unique effects once the consumer-oriented UTAUT2 constructs are included [6, 7, 9]. That habit was already the strongest predictor in a sample in which 81% of students were novices is consistent with the argument that repeated behavior gradually displaces deliberative antecedents [3].

The more consequential finding concerns necessity. In these data the strongest unique predictor of intention, habit, is at most a negligible necessary condition, while effort expectancy, which has no detectable unique effect, is among the clearest necessary conditions on the observed scale. The inversion is established on behavioral intention alone, so it does not depend on the use-behavior outcome, whose overlap with intention (HTMT 0.920) we treat as a limitation, and it holds on zero-order correlations as well as on path coefficients. We qualify it immediately: the necessity magnitudes are entangled with the predictors' Likert distributions and attenuate under a rank transformation, so the individual necessity labels are distribution-conditioned, and facilitating conditions, not effort expectancy, is the most scoring-stable necessary condition. The robust claim is the inverted ordering at the extremes. Ease of use thus behaves like a threshold condition, whose minimum level appears necessary for high intention but whose further increases add little unique prediction, whereas habit behaves like an incremental predictor: higher habit goes with higher intention, yet students without an established routine can still report high intention through other routes. The additive formulation of UTAUT2 cannot express this asymmetry, and the SEM-ANN pipeline, which shares the same additive logic, inherits that limitation.

Our necessity results partly replicate prior NCA evidence on ChatGPT acceptance [20]. Performance expectancy is again necessary, with the distributional caveat above, whereas habit is clearly not a meaningful necessary condition in a Philippine sample surveyed within a year of ChatGPT's release, when habit levels were still low and dispersed. One explanation is developmental: the necessity of a routinization construct may emerge only after a technology matures within a population. Our size-matched comparison of novices and experienced users found no evidence of such a shift, but self-rated experience is a coarse proxy that cannot separate development from selection, so the explanation remains untested and requires longitudinal or cross-cohort data.

Methodologically, the triangulation supports a sharper verdict than the customary assertion that hybrid rankings are aligned. The three attribution logics agree on which determinants matter most and least, but the agreement carries wide sampling uncertainty, and the middle of the ranking depends on the method in exactly the tier from which incremental advice is usually drawn. Three recommendations follow for hybrid acceptance research: quantify ranking agreement with resampling intervals rather than visual inspection; do not interpret mid-rank differences unless they replicate across attribution logics; and, when one ranking must be chosen, prefer an attribution with axiomatic guarantees to network-specific normalized importance, which in our data promoted price value into an actionable tier that SHAP does not support and the

structural model supports only weakly ($\beta = 0.079$, not significant after Holm correction). The predictive results carry a further message for the UTAUT2 literature. The structural model comfortably beats a naive benchmark but is out-predicted by ordinary linear regressions on use behavior, so explanatory R^2 says little about predictive usefulness [31], and we join earlier calls for PLSpredict and CVPAT to become standard reporting [16, 17]. Transparency about measurement decisions matters for the same reason: one further conventional item deletion would have manufactured a discriminant validity failure while nominally improving convergent validity.

For institutions, these implications are design hypotheses rather than procurement targets, because they rest on cross-sectional, single-country data and on distribution-conditioned necessity estimates. Read in that spirit, the results suggest a two-track logic. The first track secures the necessary conditions. In this sample few students with low perceived ease of use, weak facilitating conditions or low perceived usefulness reported high intention, so guaranteeing baseline levels through structured onboarding to prompt-based work, integration of AI assistants into existing learning platforms, reliable device and connectivity access and concrete demonstrations of usefulness in coursework is a plausible priority. The second track strengthens the engines: once these necessary conditions are in place, the largest average gains in intention plausibly come from habit formation, for example through recurring course activities that make productive ChatGPT use routine. If the necessity pattern holds and is causal, it implies securing necessary conditions before investing in engines, because engine-first investment would reach students who are already enabled while leaving constrained students below the ceiling that limits them.

Several limitations bound these conclusions. The data are cross-sectional self-reports, so neither the structural paths nor the necessity ceilings identify causal effects, and out-of-sample prediction here means generalization to unseen respondents rather than forecasting. As a secondary analysis, the study could not add marker variables or verify field procedures, and common method variance was assessed only with post hoc diagnostics. Intention and use come from the same survey and their HTMT of 0.920 exceeds the 0.90 threshold; although the bootstrap upper bound stays below 1, the two constructs overlap, which is why the intention results are primary and why objective usage logs would strengthen future work. Item purification used the full sample, including the cases later held out for the machine learning test, although no tuning decision used the test partition. The machine learning stage uses construct scores and discards item-level information. Necessity estimates on five-point Likert data are entangled with the marginal distributions, a problem that using CR-FDH as the primary ceiling and reporting the robustness analyses mitigates but does not remove. Finally, the sample represents Philippine higher education in late 2023, an early phase of generative AI diffusion, and the habit results in particular may be specific to that phase.

In conclusion, replacing the heuristic importance stage of the SEM-ANN pipeline with a framework that combines PLS-SEM, necessary condition analysis and Shapley-based machine learning, together with genuine out-of-sample evaluation, preserved the established sufficiency account of ChatGPT adoption and added what that account cannot see: the strongest predictor of intention is not a meaningful precondition,

and the clearest apparent preconditions are not the strongest predictors. Because the importance rankings agree at their extremes and the structural model offers no predictive gain over a linear benchmark, the necessity analysis supplies the complementary information that the conventional pipeline lacks. Applying the same framework to longitudinal and multi-country open datasets is the natural next step.

Methods

Data source and participants

We analyzed the open dataset of Himang et al. [21], released on Mendeley Data under a Creative Commons Attribution 4.0 license. According to the dataset description and the creators' publication [22], the original team surveyed students at 95 higher education institutions across Luzon, Visayas and Mindanao between 22 July and 29 October 2023 with an online questionnaire, collected 2,254 responses, and obtained 1,242 valid responses after data cleansing. All 1,242 respondents reported having used ChatGPT for academic work in the previous six months, the analysis-ready sample spans 78 of the 95 institutions, and it contains no missing item responses. The instrument measures the seven UTAUT2 predictors, namely performance expectancy (PE), effort expectancy (EE), social influence (SI), facilitating conditions (FC), hedonic motivation (HM), price value (PV) and habit (HAB), together with behavioral intention (BI) and use behavior (UB), each with five items on a five-point Likert scale, plus age band, gender and self-rated ChatGPT experience. The codebook labels the experience categories by weekly hours of use; because the upper bands are implausible as literal weekly hours, we used the experience codes only as ordered categories, as a binary contrast between the lowest band (novice, $n = 1{,}009$) and all higher bands (experienced, $n = 233$) in the stratified analysis and as standardized band codes in the moderation analysis.

Ethics declarations

This study is a secondary analysis of a publicly available, de-identified dataset released under an open license that permits reuse. The authors had no contact with participants, collected no new data and had no access to identifying information, and the dataset contains no participant names or contact details. Because the study involved no interaction with human participants and used only publicly available, anonymized records, no separate ethics approval or informed consent was sought for this secondary analysis. The ethics review and consent procedures of the original survey are not described in the dataset documentation. All analyses were performed in accordance with the dataset license and with relevant guidelines and regulations.

Model, hypotheses and necessity propositions

Following UTAUT2 [3], PE, EE, SI, FC, HM, PV and HAB were hypothesized to be positively associated with BI (H1 to H7), and BI, FC and HAB with UB (H8 to H10) (Supplementary Fig. S1). Age, gender and experience are candidate moderators; given

the number of possible interaction terms and the secondary nature of the data, moderation was examined in a supplementary analysis rather than as formal hypotheses. Four necessity propositions were stated separately because they concern a different relation: PE is necessary for high BI (NP1), HAB is necessary for high UB (NP2), FC is necessary for high UB (NP3), and HM and PV are not necessary for high BI (NP4). NP1 and NP2 test on independent data the necessity pattern reported for ChatGPT acceptance in a European sample [20], with the outcome separated into intention and use; NP3 reflects the physical precondition of access, devices and connectivity; and NP4 serves as a negative control showing that NCA does not mechanically declare every sufficient predictor necessary.

Stage 1: PLS-SEM

We estimated the model by PLS path modeling (Mode A reflective measurement, path weighting scheme, convergence tolerance 10^{-8} on the outer weights), implemented in Python following standard formulations [32, 33]. Measurement quality was assessed with indicator loadings, Cronbach's alpha on standardized items, composite reliability ρ_c, Dijkstra-Henseler ρ_A [34], AVE, the Fornell-Larcker criterion [35] and HTMT [29]. For HTMT values above 0.85 we computed percentile bootstrap intervals (5,000 resamples) and compared the one-sided 95% upper bound (the 95th percentile) with the threshold of 0.90 and with 1 [30]. Item purification followed common guidance [36]: the lowest-loading indicator of a construct was removed only if its loading was below 0.708, the construct's AVE was below 0.50 and at least three indicators would remain, and, as an explicit safeguard, only if the removal did not create a new HTMT value above 0.90; when several constructs qualified, the one with the lowest AVE was treated first. Path coefficients were tested with 5,000 bootstrap resamples (all of which converged), percentile 95% CIs and two-sided bootstrap P values computed from the proportion of resampled coefficients on either side of zero with an add-one correction (smallest attainable value 0.0004); because the ten hypotheses were tested jointly, we also report a Holm correction. We report effect sizes f^2 and structural variance inflation factors. Common method bias was screened with Harman's single-factor test, computed as the share of item variance explained by the first principal component, and with the full-collinearity variance inflation factor procedure with a threshold of 3.3 [37]. For the moderation analysis, each moderator (age band and experience level as standardized band codes, and gender restricted to male and female respondents, $n = 1{,}219$) was standardized and its products with the predictors were added to ordinary least squares regressions on the latent variable scores, with HC3 heteroskedasticity-consistent standard errors and two-sided t tests. The 30 interaction terms were judged against a Bonferroni-corrected level of $\alpha = 0.05/30 \approx 0.0017$.

Stage 2: necessary condition analysis

Using the standardized latent variable scores from Stage 1 as primary data and construct-mean scores as a robustness check, we estimated two ceiling lines for each predictor-outcome pair [14]: the step-wise ceiling envelopment with free disposal hull (CE-FDH) and the ceiling regression line (CR-FDH), an ordinary least squares line

through the upper-left corner points (peers) of the CE-FDH step function. The necessity effect size d is the area of the empty ceiling zone above the line divided by the scope, the rectangle spanned by the observed minima and maxima of predictor and outcome. Following NCA guidance that the smooth CR-FDH line suits continuous scores whereas the step-wise CE-FDH line suits discrete data with few levels [14, 15], we treat CR-FDH as primary; CE-FDH estimates were larger than CR-FDH estimates for every full-sample pair, so we report them as an approximate upper bound. Significance was assessed with the approximate permutation test [18]: outcome values were permuted 10,000 times and P was the add-one proportion of permuted effect sizes at least as large as the observed one (one-sided; smallest attainable value 0.0001). Percentile 95% CIs were obtained from 2,000 bootstrap resamples. Because each resample contains only a subset of the observed cases, its ceiling lies on or below the full-sample ceiling, so resampled effect sizes tend to exceed the point estimate; the intervals are therefore reported as an upward sensitivity bound rather than as a symmetric test. A predictor was considered a necessary condition when $d \geq 0.10$ (a medium effect [14]), $P < 0.05$ and the relation was theoretically plausible [15]. CE-FDH bottleneck tables report, for outcome levels from 0 to 100% of the observed range, the minimum percentage of the observed predictor range required.

Robustness analyses and experience strata

For each predictor of intention we computed the mean, standard deviation and skewness of the construct-mean score and the share of respondents whose construct mean was at least 4.75, that is, with at most one item one point below the scale maximum, a definition that is comparable for four- and five-item scales. We computed Spearman correlations across the seven predictors between the CR-FDH effect size and the predictor mean and skewness, with exact two-sided permutation P values over all 5,040 orderings, and Spearman correlations between the sufficiency ranking (absolute path coefficients, or zero-order correlations with BI) and the necessity ranking. We re-estimated NCA after transforming the construct-mean scores of each predictor and of the outcome to uniform marginals with the rank transform $(r - 0.5)/n$, which removes differences in distribution shape and range restriction, and tested the rank-based effects with 999 permutations. Three further checks are reported in Supplementary Table S13: the range of CR-FDH effect sizes after deleting each ceiling-defining case (peer) in turn; the share of test-partition cases lying above CR-FDH ceilings estimated on the development partition, using the partitions of Stage 3; and the path coefficients and effect sizes obtained when BI1 is retained. The experience-stratified analysis re-estimated CE-FDH and CR-FDH effect sizes with 10,000-permutation tests on construct-mean scores separately for novices and experienced users. Because FDH-based effect sizes depend on sample size, the experienced group was compared with the distribution of estimates from 1,000 random novice subsamples of the same size ($n = 233$, drawn without replacement).

Stage 3: machine learning and Shapley attributions

With construct-mean scores as features, we predicted BI from the seven predictors and UB from BI, FC and HAB, mirroring the structural model. The data were split once at random into a development partition (85%, $n = 1{,}055$) and a test partition (15%, $n = 187$) that remained locked until final evaluation. XGBoost [23] hyperparameters were tuned by random search (200 draws from a grid of 46,656 configurations, of which 199 were unique) over the number of trees {100, 200, 400, 800}, maximum depth {2, 3, 4, 6}, learning rate {0.01, 0.03, 0.05, 0.1}, row and column subsampling {0.6, 0.8, 1.0}, minimum child weight {1, 5, 10}, split-loss penalty γ {0, 0.1, 0.5}, L1 penalty {0, 0.1, 1} and L2 penalty {1, 5, 10}, selecting the configuration with the lowest five-fold cross-validated RMSE on the development partition. The ANN comparator, representing the conventional SEM-ANN choice, was a single-hidden-layer multilayer perceptron with rectified linear units, trained with the Adam optimizer on standardized inputs, with 2, 4, 8 or 16 hidden units, an L2 penalty in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, an initial learning rate in $\{10^{-3}, 10^{-2}\}$ and early stopping on a 10% validation split (patience 30 epochs, at most 4,000 epochs); it was tuned by a 24-draw random search over the 32 configurations under the same five-fold scheme. Ordinary least squares regression served as the linear comparator. Because the cross-validated RMSE of each tuned model is the score of the selected configuration, without nested cross-validation, it is optimistically biased relative to the untuned linear model. Test performance is reported as RMSE, mean absolute error (MAE) and R^2, and test-set differences in RMSE from linear regression were assessed with a paired bootstrap of the 187 test cases (10,000 resamples) and a paired t test on squared errors. Global SHAP importance is the mean absolute TreeSHAP value [11, 12] on the test partition, and SHAP interaction values were computed on the full sample. ANN normalized importance followed Garson's weight-partitioning algorithm [38], a weight-based alternative to the sensitivity-based index common in SEM-ANN studies.

Agreement among the three importance rankings for BI (mean absolute SHAP value, absolute path coefficient and ANN normalized importance) was tested with Kendall's coefficient of concordance W against a null distribution obtained from 10,000 independent random permutations of each method's ranks (one-sided), because the chi-square approximation is unreliable with seven items and three raters. Pairwise agreement was summarized with Spearman's ρ and Kendall's τ between the three importance vectors as estimated above (SHAP on the test partition, path coefficients on the full sample and ANN importance from the network fitted to the development partition). Percentile 95% CIs came from 500 bootstrap resamples in which, on each resample, the PLS-SEM model and XGBoost and ANN models with the tuned hyperparameters held fixed were refitted to the whole resample and SHAP values were computed in-sample; because this differs from the computation of the point estimates, the intervals are not centered on them.

Stage 4: out-of-sample assessment of the structural model

PLSpredict [16] was run with ten folds and ten repetitions. In each fold the model was estimated on the training cases, and the endogenous indicators of the held-out

cases were predicted from their exogenous indicators through the estimated weights, path coefficients and loadings, with predicted intention scores feeding the prediction of use behavior (the earliest-antecedents approach). Predictions were compared with an indicator-average (IA) benchmark, the training mean, and with a linear-model (LM) benchmark, an ordinary least squares regression of each endogenous indicator on all exogenous indicators, using Q^2_{predict}, RMSE and MAE averaged over repetitions. CVPAT [17, 24] compared the per-case squared prediction losses of PLS-SEM and each benchmark, averaged over the indicators of a construct (or of all endogenous constructs for the overall test), with one-sample t tests on the 1,242 per-case loss differences; negative average loss differences favor PLS-SEM. We used two-sided tests, rather than the one-sided test of PLS-SEM superiority in the original CVPAT [24], because the benchmark can also predict better.

Statistical reporting and software

All tests used $\alpha = 0.05$ and were two-sided, except the NCA and concordance permutation tests, which are one-sided by construction. Resampling P values below 0.001 are reported as $P < 0.001$ (the smallest attainable values are 0.0004 for 5,000 bootstrap resamples and 0.0001 for 10,000 permutations); exact parametric P values for CVPAT are given in Table 4. A fixed random seed (42) was used throughout. Analyses were run in Python 3.10.8 with pandas 2.3.3, NumPy 1.26.4, scikit-learn 1.6.1, XGBoost 2.1.3, SHAP 0.46.0, SciPy 1.13.1 and Matplotlib 3.9.2. The PLS-SEM, NCA, PLSpredict and CVPAT routines were implemented from their primary methodological sources and checked with unit tests on hand-computed cases, and rerunning the full pipeline from the raw data reproduced all reported values.

Use of large language models

An AI assistant (Claude, Anthropic) was used to assist with writing and reviewing the analysis code and with copy-editing the manuscript. All code, analyses and text were checked by the authors, who take full responsibility for the content of this article.

Data availability

The dataset analyzed in this study is openly available in Mendeley Data at https://doi.org/10.17632/7fz5rkz6bw.1 [21]. All derived results are available in the code repository listed below.

Code availability

The analysis code, together with all derived results, tables and figures, is openly available at https://github.com/AnhNguyenVLU/ChatGPT_SEM_NCA_XAI.

References

[1] Digital Education Council. Global AI student survey 2024. https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024 (2024).

[2] Freeman, J. Student generative AI survey 2025. HEPI Policy Note 61, Higher Education Policy Institute https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/ (2025).

[3] Venkatesh, V., Thong, J. Y. L. & Xu, X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Q.* **36**, 157–178 (2012).

[4] Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. User acceptance of information technology: toward a unified view. *MIS Q.* **27**, 425–478 (2003).

[5] Leong, L.-Y., Hew, T.-S., Ooi, K.-B., Tan, G. W.-H. & Koohang, A. An SEM-ANN approach: guidelines in information systems research. *J. Comput. Inf. Syst.* **65**, 706–737 (2025).

[6] Strzelecki, A. To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interact. Learn. Environ.* **32**, 5142–5155 (2024).

[7] Habibi, A. *et al.* ChatGPT in higher education learning: acceptance and use. *Comput. Educ. Artif. Intell.* **5**, 100190; https://doi.org/10.1016/j.caeai.2023.100190 (2023).

[8] Sobaih, A. E. E., Elshaer, I. A. & Hasanein, A. M. Examining students' acceptance and use of ChatGPT in Saudi Arabian higher education. *Eur. J. Investig. Health Psychol. Educ.* **14**, 709–721 (2024).

[9] Menon, D. & Shilpa, K. "Chatting with ChatGPT": analyzing the factors influencing users' intention to use the Open AI's ChatGPT using the UTAUT model. *Heliyon* **9**, e20962; https://doi.org/10.1016/j.heliyon.2023.e20962 (2023).

[10] Foroughi, B. *et al.* Determinants of intention to use ChatGPT for educational purposes: findings from PLS-SEM and fsQCA. *Int. J. Hum. Comput. Interact.* **40**, 4501–4520 (2024).

[11] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).

[12] Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

[13] de Andrés-Sánchez, J., Arias-Oliva, M., Souto-Romero, M. & Llorens-Marín, M. Integrating machine learning techniques and the unified theory of acceptance and use of technology to evaluate drivers for the acceptance of blockchain-based loyalty programmes. *Comput. Econ.* https://doi.org/10.1007/s10614-025-11270-y (2026).

[14] Dul, J. Necessary condition analysis (NCA): logic and methodology of "necessary but not sufficient" causality. *Organ. Res. Methods* **19**, 10–52 (2016).

[15] Richter, N. F., Schubring, S., Hauff, S., Ringle, C. M. & Sarstedt, M. When predictors of outcomes are necessary: guidelines for the combined use of PLS-SEM and NCA. *Ind. Manag. Data Syst.* **120**, 2243–2267 (2020).

[16] Shmueli, G. *et al.* Predictive model assessment in PLS-SEM: guidelines for using PLSpredict. *Eur. J. Mark.* **53**, 2322–2347 (2019).

[17] Sharma, P. N., Liengaard, B. D., Hair, J. F., Sarstedt, M. & Ringle, C. M. Predictive model assessment and selection in composite-based modeling using PLS-SEM: extensions and guidelines for using CVPAT. *Eur. J. Mark.* **57**, 1662–1677 (2023).

[18] Dul, J., van der Laan, E. & Kuik, R. A statistical significance test for necessary condition analysis. *Organ. Res. Methods* **23**, 385–395 (2020).

[19] Hauff, S., Richter, N. F., Sarstedt, M. & Ringle, C. M. Importance and performance in PLS-SEM and NCA: introducing the combined importance-performance map analysis (cIPMA). *J. Retail. Consum. Serv.* **78**, 103723; https://doi.org/10.1016/j.jretconser.2024.103723 (2024).

[20] Kayser, I. & Gradtke, M. Unlocking AI acceptance: an integration of NCA and PLS-SEM to analyse the acceptance of ChatGPT. *J. Decis. Syst.* https://doi.org/10.1080/12460125.2024.2443231 (2024).

[21] Himang, C. M. *et al.* Acceptance and use of ChatGPT among college students: a dataset from select higher education institutions in Luzon, Visayas, and Mindanao. *Mendeley Data* https://doi.org/10.17632/7fz5rkz6bw.1 (2023).

[22] Himang, C. M. *et al.* Adoption of ChatGPT among higher education students. *Int. J. Inf. Commun. Technol. Educ.* **22**, 1–29 (2026).

[23] Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).

[24] Liengaard, B. D. *et al.* Prediction: coveted, yet forsaken? Introducing a cross-validated predictive ability test in partial least squares path modeling. *Decis. Sci.* **52**, 362–392 (2021).

[25] Guo, H. & Erdenebold, T. Factors influencing intention to adopt an AI chatbot for learning in higher education: an integrated PLS-SEM, IPMA, and ANN approach. *Comput. Educ. Artif. Intell.* **9**, 100477; https://doi.org/10.1016/j.caeai.2025.100477 (2025).

[26] Zeng, D., Xu, X., Zhu, T., Li, Y. & Li, Q. A two-step structural equation modeling and explainable machine learning framework for understanding university students' adoption of generative AI: balancing intrinsic motivations and perceived risks. *Front. Psychol.* **17**, 1743722; https://doi.org/10.3389/fpsyg.2026.1743722 (2026).

[27] Leong, L.-Y., Hew, T.-S., Ooi, K.-B. & Chau, P. Y. K. "To share or not to share?" A hybrid SEM-ANN-NCA study of the enablers and enhancers for mobile sharing economy. *Decis. Support Syst.* **180**, 114185; https://doi.org/10.1016/j.dss.2024.114185 (2024).

[28] Aziz, F., Li, C. & Khan, A. U. Transforming behavioral intention and academic performance: ChatGPT-4.0 insights through SEM, ANN, and cIPMA analysis. *Inf. Dev.* **41**, 933–956 (2025).

[29] Henseler, J., Ringle, C. M. & Sarstedt, M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* **43**, 115–135 (2015).

[30] Franke, G. & Sarstedt, M. Heuristics versus statistics in discriminant validity testing: a comparison of four procedures. *Internet Res.* **29**, 430–447 (2019).

[31] Chin, W. *et al.* Demystifying the role of causal-predictive modeling using partial least squares structural equation modeling in information systems research. *Ind. Manag. Data Syst.* **120**, 2161–2209 (2020).

[32] Lohmöller, J.-B. *Latent Variable Path Modeling with Partial Least Squares* (Physica-Verlag, 1989).

[33] Hair, J. F., Hult, G. T. M., Ringle, C. M. & Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* 3rd edn (Sage, 2022).

[34] Dijkstra, T. K. & Henseler, J. Consistent partial least squares path modeling. *MIS Q.* **39**, 297–316 (2015).

[35] Fornell, C. & Larcker, D. F. Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* **18**, 39–50 (1981).

[36] Hair, J. F., Risher, J. J., Sarstedt, M. & Ringle, C. M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **31**, 2–24 (2019).

[37] Kock, N. Common method bias in PLS-SEM: a full collinearity assessment approach. *Int. J. e-Collab.* **11**, 1–10 (2015).

[38] Garson, G. D. Interpreting neural-network connection weights. *AI Expert* **6**, 46–51 (1991).

Acknowledgements

We thank the creators of the dataset for making their item-level data publicly available.

Author contributions

N.H.L.: conceptualization, methodology, software, formal analysis, data curation, visualization and writing (original draft). N.T.A.: conceptualization, validation, investigation, writing (review and editing), supervision and project administration. Both authors reviewed and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Additional information

Competing interests. The authors declare no competing interests.

Supplementary information accompanies this paper as a separate file.

Correspondence and requests for materials should be addressed to N.T.A.

Figures

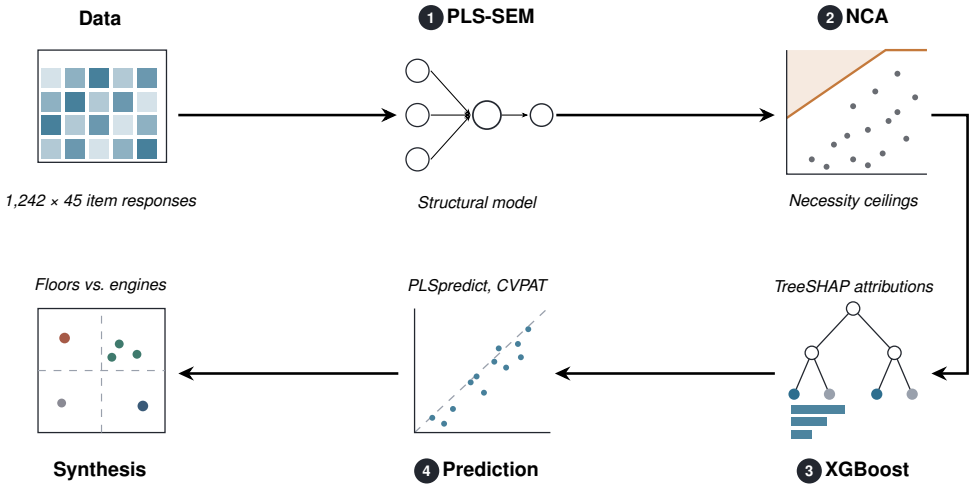


Fig. 1 Four-stage analytical framework. From the item-level survey data (1,242 respondents, 45 items), stage 1 estimates and validates the structural model with PLS-SEM; stage 2 tests necessity with ceiling lines, where an empty upper-left region of the predictor-outcome scatter indicates a necessary condition; stage 3 fits a tuned XGBoost model interpreted with TreeSHAP and benchmarked against an artificial neural network and linear regression on a locked test set; and stage 4 assesses out-of-sample prediction of the structural model with PLSpredict and CVPAT. The synthesis places each predictor by sufficiency and necessity, using the colors of Fig. 3: floors (terracotta), core preconditions (green), engines (dark blue) and peripheral predictors (grey).

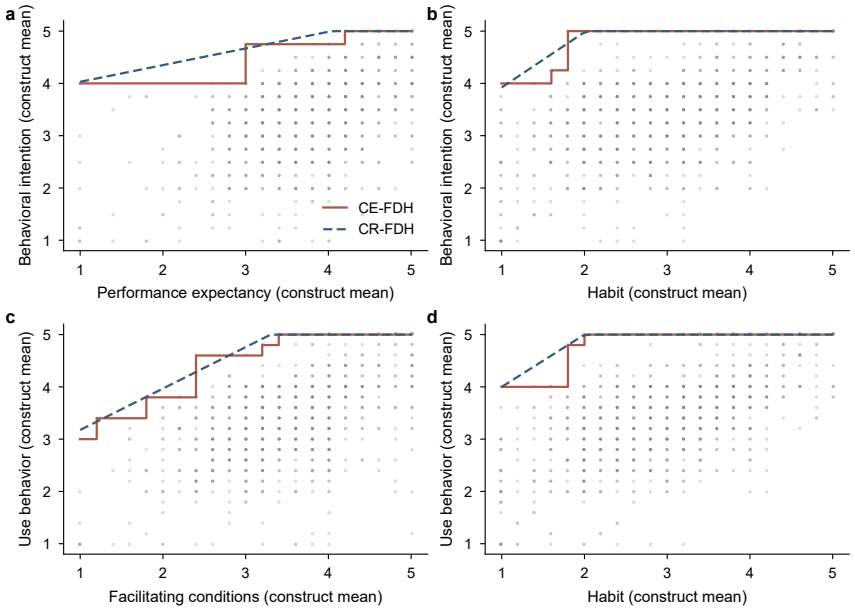


Fig. 2 Necessity ceilings for four predictor-outcome pairs. a, Performance expectancy and behavioral intention. **b**, Habit and behavioral intention. **c**, Facilitating conditions and use behavior. **d**, Habit and use behavior. Points are construct-mean scores (scale 1 to 5) of the 1,242 respondents; overlapping points appear darker. Solid lines show the ceiling envelopment (CE-FDH) and dashed lines the ceiling regression (CR-FDH). The empty upper-left regions in **a** and **c** show a ceiling constraint (for performance expectancy it is borderline on these construct-mean scores, CR-FDH $d = 0.092$; Table 3), whereas the ceilings for habit in **b** and **d** reach the scale maximum almost immediately.

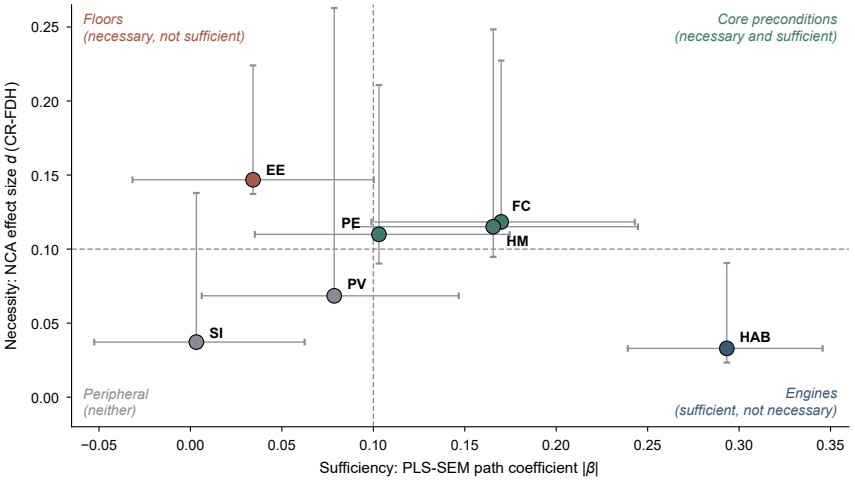


Fig. 3 Necessity-sufficiency map for the seven predictors of behavioral intention. The horizontal axis shows sufficiency (absolute PLS-SEM path coefficient) and the vertical axis necessity (CR-FDH effect size d on latent variable scores). Error bars are percentile 95% bootstrap confidence intervals (5,000 resamples for path coefficients, 2,000 for d). Horizontal intervals are those of the signed coefficients, all of whose point estimates are positive, so intervals that include zero extend to the left of the origin; vertical intervals are shifted upward because each resample contains only a subset of the cases. Dashed lines mark the thresholds of 0.10. Marker colors match the quadrant labels, with necessity requiring $d \geq 0.10$ and permutation $P < 0.05$, and sufficiency requiring $|\beta| \geq 0.10$. PE, performance expectancy; EE, effort expectancy; SI, social influence; FC, facilitating conditions; HM, hedonic motivation; PV, price value; HAB, habit.

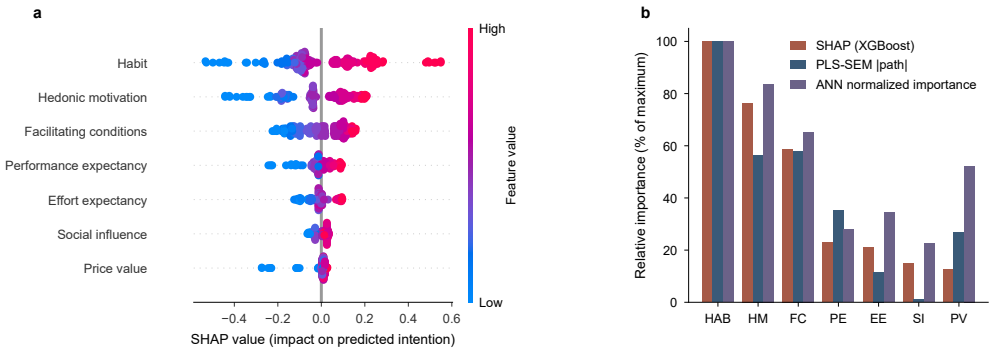


Fig. 4 Importance attributions for behavioral intention. a, SHAP summary for the tuned XGBoost model on the locked test set ($n = 187$); each point is one student, jittered vertically, the horizontal position is the contribution to predicted intention, and the color is the student's construct-mean score on that predictor, from low (blue) to high (red). **b**, Relative importance under three attribution methods, each normalized to its maximum: mean absolute SHAP value on the test set ($n = 187$), absolute PLS-SEM path coefficient on the full sample ($n = 1{,}242$) and ANN normalized importance from Garson's algorithm for the network fitted to the development partition ($n = 1{,}055$). Bars are point estimates; their uncertainty is summarized in Supplementary Table S11. Abbreviations as in Fig. 3.

Tables

Table 1 Structural model estimates (n = 1,242; 5,000 bootstrap resamples).

H	Path	β	SE	t	P	95% CI	f^2	VIF	Supported
H1	PE → BI	0.103**	0.036	2.89	0.003	[0.035, 0.175]	0.009	2.16	Yes
H2	EE → BI	0.034	0.034	1.02	0.300	[−0.032, 0.100]	0.001	2.19	No
H3	SI → BI	0.003	0.029	0.11	0.866	[−0.053, 0.062]	0.000	1.75	No
H4	FC → BI	0.170***	0.037	4.65	<0.001	[0.099, 0.243]	0.024	2.21	Yes
H5	HM → BI	0.166***	0.040	4.15	<0.001	[0.089, 0.245]	0.019	2.57	Yes
H6	PV → BI	0.079*	0.036	2.19	0.034	[0.006, 0.147]	0.006	1.96	Yes
H7	HAB → BI	0.293***	0.027	10.92	<0.001	[0.239, 0.346]	0.105	1.48	Yes
H8	BI → UB	0.452***	0.028	16.32	<0.001	[0.398, 0.505]	0.299	1.66	Yes
H9	FC → UB	0.176***	0.027	6.42	<0.001	[0.124, 0.232]	0.052	1.46	Yes
H10	HAB → UB	0.289***	0.025	11.67	<0.001	[0.242, 0.339]	0.138	1.48	Yes

β, standardized path coefficient; SE, bootstrap standard error; $t = |\beta|/\mathrm{SE}$, shown for reference; P, two-sided bootstrap P value (smallest attainable value 0.0004), on which Supported ($\alpha = 0.05$) is based; under a Holm correction across the ten paths, H6 is not supported; CI, percentile confidence interval; f^2, effect size; VIF, structural variance inflation factor. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. R^2 is 0.446 for behavioral intention (BI) and 0.589 for use behavior (UB). PE, performance expectancy; EE, effort expectancy; SI, social influence; FC, facilitating conditions; HM, hedonic motivation; PV, price value; HAB, habit.

Table 2 Necessity effect sizes for behavioral intention and use behavior (CR-FDH primary, CE-FDH as upper bound).

	CR-FDH (primary) latent variable scores			CE-FDH latent variable scores		CE-FDH construct means
Path	d	95% CI	P	d	P	d
PE → BI	0.110	[0.090, 0.211]	0.002	0.139	<0.001	0.144
EE → BI	0.147	[0.137, 0.224]	<0.001	0.160	<0.001	0.156
SI → BI	0.037	[0.037, 0.138]	0.205	0.074	0.005	0.062
FC → BI	0.118	[0.118, 0.227]	<0.001	0.137	<0.001	0.141
HM → BI	0.115	[0.095, 0.248]	<0.001	0.132	<0.001	0.134
PV → BI	0.068	[0.068, 0.263]	0.087	0.107	0.004	0.116
HAB → BI	0.033	[0.023, 0.091]	0.002	0.044	<0.001	0.047
BI → UB	0.187	[0.176, 0.218]	<0.001	0.199	<0.001	0.200
FC → UB	0.135	[0.134, 0.214]	<0.001	0.151	<0.001	0.153
HAB → UB	0.042	[0.042, 0.099]	<0.001	0.052	<0.001	0.053

d, necessity effect size ($d \geq 0.10$ medium, $d \geq 0.30$ large). CR-FDH, ceiling regression; CE-FDH, ceiling envelopment, which gave larger values for every pair and is reported as an approximate upper bound. CIs are percentile intervals from 2,000 bootstrap resamples; because each resample contains only a subset of the cases, they are shifted upward and their lower bound is often close to the point estimate. P values are one-sided, from 10,000-permutation approximate tests (smallest attainable value 0.0001). The final column is the CE-FDH estimate on construct-mean scores. Abbreviations as in Table 1.

Table 3 Distributional robustness of the necessity estimates for behavioral intention.

Predictor	Mean	SD	Skew	% near max.	d CR-FDH (latent)	d CR-FDH (means)	d CR-FDH (ranks)	$\lvert\beta\rvert$
Performance expectancy (PE)	3.77	0.65	−0.83	7.1	0.110	0.092	0.043	0.103
Effort expectancy (EE)	3.86	0.61	−0.82	9.9	0.147	0.144	0.036	0.034
Social influence (SI)	3.34	0.72	−0.35	2.9	0.037	0.031	0.004	0.003
Facilitating conditions (FC)	3.52	0.73	−0.38	5.9	0.118	0.119	0.055	0.170
Hedonic motivation (HM)	3.69	0.66	−0.67	6.5	0.115	0.097	0.018	0.166
Price value (PV)	3.62	0.66	−0.45	5.8	0.068	0.074	0.025	0.079
Habit (HAB)	3.02	0.88	−0.18	2.1	0.033	0.034	0.007	0.293

Mean, SD and skewness (Skew) are for construct-mean scores (n = 1,242). % near max. is the share of respondents with a construct mean of at least 4.75 (at most one item one point below the scale maximum). The d columns are CR-FDH necessity effect sizes on latent variable scores, construct-mean scores and rank-transformed (uniform-marginal) construct-mean scores. Across the seven predictors, d (latent) tends to increase with the mean (Spearman $\rho = 0.75$, exact two-sided permutation $P = 0.066$) and falls below 0.10 for every predictor under the rank transformation (0.004 to 0.055), indicating that necessity magnitudes are entangled with the marginal distributions.

Table 4 Out-of-sample predictive assessment. a, CVPAT average loss differences of PLS-SEM relative to two benchmarks. **b**, Machine learning models on the development and locked test partitions.

a					
Construct	Benchmark	$\overline{D}$	t	P	
BI	Indicator average	−0.2435	−34.72	$< 10^{-15}$	
BI	Linear model	−0.0017	−0.80	0.424	
UB	Indicator average	−0.2792	−40.90	$< 10^{-15}$	
UB	Linear model	0.0184	8.29	$< 10^{-15}$	
Overall	Indicator average	−0.2633	−40.99	$< 10^{-15}$	
Overall	Linear model	0.0095	5.84	6.6×10^{-9}	
b					
Target	Model	CV RMSE	Test RMSE	Test MAE	Test R^2
BI	XGBoost	0.5552	0.5730	0.4579	0.3924
BI	ANN (single hidden layer)	0.5526	0.5909	0.4707	0.3539
BI	Linear regression	0.5521	0.5889	0.4691	0.3583
UB	XGBoost	0.4845	0.4875	0.3676	0.5168
UB	ANN (single hidden layer)	0.4833	0.4834	0.3702	0.5249
UB	Linear regression	0.4802	0.4840	0.3703	0.5237

a, $\overline{D}$ is the mean per-case difference in squared prediction loss (PLS-SEM minus benchmark) from PLSpredict with ten folds and ten repetitions; negative values favor PLS-SEM. P values are from two-sided one-sample t tests on the 1,242 per-case differences. **b**, CV RMSE is the five-fold cross-validated root mean squared error on the development partition (n = 1,055); for XGBoost and the ANN it is the score of the selected configuration and is therefore optimistically biased. Test metrics use the locked test partition (n = 187); the paired bootstrap 95% CI of the test RMSE difference between XGBoost and linear regression for BI is −0.032 to 0.001. BI, behavioral intention; UB, use behavior; ANN, artificial neural network; MAE, mean absolute error.